DOCUMENT RESUME

ED 374 162 TM 022 091

AUTHOR Wheeler, Patricia H.
TITLE The Use of Confidence Intervals When Interpreting Test Scores. EREAPA Publication Series No. 93-4.
INSTITUTION EREAPA Associates, Livermore, CA.
PUB DATE 93
NOTE 9p.
AVAILABLE FROM EREAPA Associates, 2840 Waverly Way, Livermore, CA 94550-1740 ($3).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cutting Scores; *Error of Measurement; *Scores; *Statistical Distributions; *Test Interpretation; Test Results; Test Theory; *True Scores
IDENTIFIERS *Confidence Intervals (Statistics)

ABSTRACT

A person's obtained score on a test provides an estimate of the individual's "true" score on that test. The obtained score is considered to have two parts, the true component and the error component. Classical test theory assumes that obtained scores for an individual over multiple administrations of the same test will lie symmetrically around the individual's true score. Confidence intervals can be used to determine the range within which the true score is apt to fall and to identify a second critical score when cut scores are used. To determine the upper and lower limits of a confidence interval, the standard error of measurement is multiplied by a number depending on the level of confidence needed for the situation. Multipliers come from a normal distribution based on the percentage of the total area under the normal curve between various standard deviations above and below the mean. How confidence intervals can be used to assess change, and when they should and should not be used, are discussed. (Contains 7 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 374 162

EREAPA ASSOCIATES

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Patricia H. Wheeler
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EREAPA Publication Series No. 93-4

The Use of Confidence Intervals When Interpreting Test Scores

Patricia H. Wheeler, M.B.A., Ph.D.

TM042091

BEST COPY AVAILABLE

EREAPA
Publication
Series
No. 93-4

The Use of Confidence Intervals When Interpreting Test Scores

Patricia H. Wheeler, M.B.A., Ph.D.

EREAPA Associates
2840 Waverley Way
Livermore, California 94550-1740

1993

Copyright © 1993 by Patricia H. Wheeler.

The author expresses her appreciation to John Bianchini, Evelyn Hawkins, and Paul K, Wheeler for their comments and suggested changes on an earlier version of this paper.

The Use of Confidence Intervals When Interpreting Test Scores

Patricia H. Wheeler, M.B.A., Ph.D.
EREAPA Associates
Livermore, California

A person's obtained score on a test provides an estimate of an individual's "true" score on that test. The obtained score is considered to have two parts: the "true" component and the error component (Gulliksen, 1950). The "true" component is a hypothetical score that represents an assessment result which is entirely free of error. It can be thought of as the average score of an infinite series of the same (or an equivalent) test, but without practice effect or any change in the person taking the tests across the series of administrations. "The error score is that part of the obtained score which is unsystematic, random, and due to chance. It is the accumulated effects of all uncontrolled and unspecified influencing factors included in the test score." (Harvill, 1991, p. 33) Examples of such factors that contribute to error score are lack of reliability in the test, variability in administrative procedures or settings across test administrations, the limited sampling of an individual's knowledge or skills in the area being tested, and interaction of such factors (Wheeler and Haertel, 1993). Classical test theory assumes that the obtained scores for an individual over multiple administrations of the same test (assuming no practice effect and no change in that individual on the attributes being tested) will lie symmetrically around that individual's "true" score, in which case the error scores will also be symmetrically distributed and will have an average of zero (Traub and Rowley, 1991).

Although it is not feasible to determine an individual's "true" score, we can use confidence intervals to determine the range, or interval, within which the true score is apt to fall. Confidence intervals allow us to make statements such as, "It is likely that this job applicant's true score falls between 148 and 156;" or "Since the confidence intervals for your student's fall and spring test scores overlap, we cannot say for certain that he has shown any improvement in his math problem solving skills."

When test users rely on specific cut scores to make decisions about individuals or to advise them on educational and career alternatives, they should consider the confidence interval (or score band). This practice will reduce the possibilities of false negatives

(failing a person whose true score is above the cut score) and false positives (passing a person whose true score is below the cut score). The consequences of false negatives are that individuals may not be given opportunities at which they could succeed, or they may be placed in a less challenging (and possibly boring) educational program or job. In the case of false positives, persons may endanger others (especially for licensing decisions), or waste time and resources in an educational program or job at which their chances of success are low.

If a program is using a cut score, a second critical score can be established, using confidence intervals. This would identify those persons whose test scores are between the cut score and this critical score (i.e., in the "uncertain" category, as described by Livingston and Zieky, 1982) and who are the potential false negatives or false positives. They might be retested with a parallel form of the test, or possibly with an easier level of the test, in order to obtain more information about that person's knowledge and skills in the areas being tested. Because of the benefit of practice effect, one might average the two scores (using scaled scores, not raw scores, since the person probably took another form or level). Or one might give the person the benefit of the doubt and go with the higher score. Such decisions should be made carefully, depending on the potential consequences of such decisions Test users should always consider additional information to support any decisions made using test scores.

Computing the Limits of the Confidence Interval

To determine the upper and lower limits of a confidence interval, we multiply the standard error of measurement (SEM) by a number, depending upon the level of confidence needed for the situation. The SEM is the estimated standard deviation of the distribution of the error scores (i.e., the differences between the obtained score and the true scores).

The multipliers come from a normal distribution, based on the percentage of the total area under the normal curve between various standards deviations above and below the mean. For example, 68% of the total area is between one standard deviation below and one standard deviation above the mean. Therefore, a confidence level of 68% has a multiplier of 1.00. The multipliers for five levels of confidence are provided below:

Confidence Level	SEM Multiplier
50 %	0.67
68 %	1.00
75%	1.14
90 %	1.65
95 %	1.96
99 %	2.58

The resulting number is added to and subtracted from the obtained score to determine the upper and lower limits, respectively, of the confidence interval. For example, if a person had an obtained score of 32 on a test with an SEM of 2.3, for a confidence level of 90%, 2.3 is multiplied by 1.65. This results in the number 3.8, which rounds off to 4. The confidence interval in this case is defined by 32 ± 4, or 28 to 36. This means that we can infer, with 90% likelihood, that this person's true score falls within this interval. At this level of confidence, 10% of the confidence intervals computed from an individual's obtained score would not include that individual's true score (Feldt and Brennan, 1989).

Confidence intervals are computed using raw scores or scaled scores, depending on whether the SEM is expressed in raw or scaled score units. SEM values vary for scores at different points of the distribution. Therefore, test publishers sometimes provide estimated SEM values for several score levels. If this is the case, the test user should select the SEM value for the score closest to the individual's obtained score rather than the SEM for the total test. "Using the SEM calculated for the total test score range may mask or enhance some score differences depending upon where they occur within the score range" (Harvill, 1991, p. 38). For example, a case with a pretest/posttest score difference of 6 may not show overlapping confidence intervals if the scores are near the middle of the distribution, but may if those scores are either very high or very low. This practice of using different SEM values for different individuals also applies when estimated SEM values are provided for various types of examinees (e.g., college-bound versus non-college-bound students; engineering majors versus liberal arts majors).

Percentile bands are one frequently reported type of confidence interval. To determine the percentile band, computations are made with raw or scaled scores first and then the corresponding percentile ranks are found in a norms table for the upper and lower limits of the confidence interval. These two percentile ranks define the percentile band. Percentiles are not an equal interval scale, as are raw scores and scaled scores. That is, a one-point difference in percentile rank near the middle of the distribution of scores can

represent a very small difference in performance level, whereas at either end of the distribution, a one-point difference in percentile rank can represent several points difference in raw or scaled score. Therefore, the percentile rank corresponding to the individual's obtained score will not necessarily fall half way between the two percentile ranks defining the percentile band.

Confidence intervals also apply to group averages. However, when computing such intervals, one must use the SEM based on a distribution of group averages, not of individual scores. The variance of a distribution of group averages will be much smaller than that of individual scores and thus the SEM, an indicator of such variance, will also be much smaller (Wheeler and Haertel, 1993).

Using Confidence Intervals to Assess Change

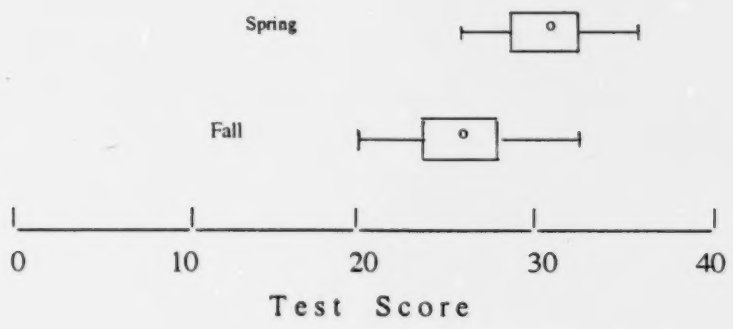
When comparing scores over time for individuals, one can use confidence intervals. If there is no overlap in the confidence interval for the first time a test was taken and the second time it was taken, we can infer a real change in the level of performance. If the confidence intervals overlap, there may not have been a real change in performance. For example, a student tested in the fall scored 26 on a test and, in the spring on an equivalent test, scored 31. The SEM for the fall test was 2.8 and for the spring test 2.6. The confidence intervals for this student's scores are computed as follows for a 95% confidence level:

Fall Test Confidence Interval	Spring Test Confidence Interval
2.8 x 1.96 = 5.48	2.6 x 1.96 = 5.10
26 ± 5.48 = 20.42 to 31.48 =	31 ± 5.10 = 25.99 to 36.10 =
20 to 32	26 to 36

We cannot say, with a high degree of confidence, that there has been a real change in performance from fall to spring since the confidence intervals overlap. If we were willing to have a 50% confidence level, then the confidence intervals would not overlap, as shown below. We could infer that there was a real difference, but with less confidence than if we had used the 95% confidence level and found no overlap.

Fall Test Confidence Interval	Spring Test Confidence Interval
2.8 x 0.67 = 1.88	2.6 x 0.67 = 1.74
26 ± 1.88 = 24.12 to 27.88 =	31 ± 1.74 = 29.26 to 32.74 =
24 to 28	29 to 32

We can use box-and-whisker diagrams to illustrate these confidence intervals graphically. In this example, the box represents the 50% confidence level, and the whiskers, the 95% confidence level.



When to Use and Not Use Confidence Intervals

Confidence interval is not the same as statistical significance and should not be interpreted in that manner (McNemar, 1962). Statistical significance refers to hypothesis testing, not to interpretation of test scores or test score gains.

In cases where test reliability is low and/or an individual's score is at the extreme end of the score distribution (very high or very low), the use of confidence intervals is not recommended. In the first case, a more reliable test should be administered. In the second case, where feasible, the individual should be given a more difficult or an easier form of the test, a level of the test that matches that individual's functional-level with regard to the attribute being tested.

Confidence intervals should be used when interpreting test scores, except as noted above. They are a reminder to those using the score as well as to the examinee that no test is a perfect measure of an attribute and that the obtained score is only an estimate of that individual's level of performance.

References

Feldt, Leonard S.; & Brennan, Robert L. (1989). Reliability. In Robert L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105-146). New York, NY: Macmillan Publishing Company.

Gulliksen, Harold. (1987). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers. (Originally published in 1950)

Harvill, Leo M. (1991, Summer). Standard error of measurement (NCME Instructional Module). *Educational Measurement: Issues and Practice*, *10*(2), 33-41.

Livingston, Samuel A.; & Zieky, Michael J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

McNemar, Quinn. (1962). *Psychological statistics* (3rd ed.). New York, NY: John Wiley and Sons, Inc.

Traub, Ross E.; & Rowley, Glenn L. (1991, Spring). Understanding reliability (NCME Instructional Module). *Educational Measurement: Issues and Practice*, *10*(1), 37-45.

Wheeler, Patricia; & Haertel, Geneva D. (1993). *A resource handbook on performance assessment and measurement: A tool for students, practitioners, and policymakers*. Berkeley, CA: The Owl Press.